An introduction to TiDB

Daniël van Eeden

HighLoad++ Armenia

Co-organizer

Yandex

Introduction

- Daniël van Eeden
- PingCAP is the company behind TiDB
  - Founded in 2015
  - Offices all around the world
  - Customers all over the world
  - Open Source Culture

Basics

- The TiDB database platform is an opensource MySQL compatible database system.
- By re-thinking and re-writing the database with scalability in mind PingCAP created a database that is easily scalable and has high availability builtin.

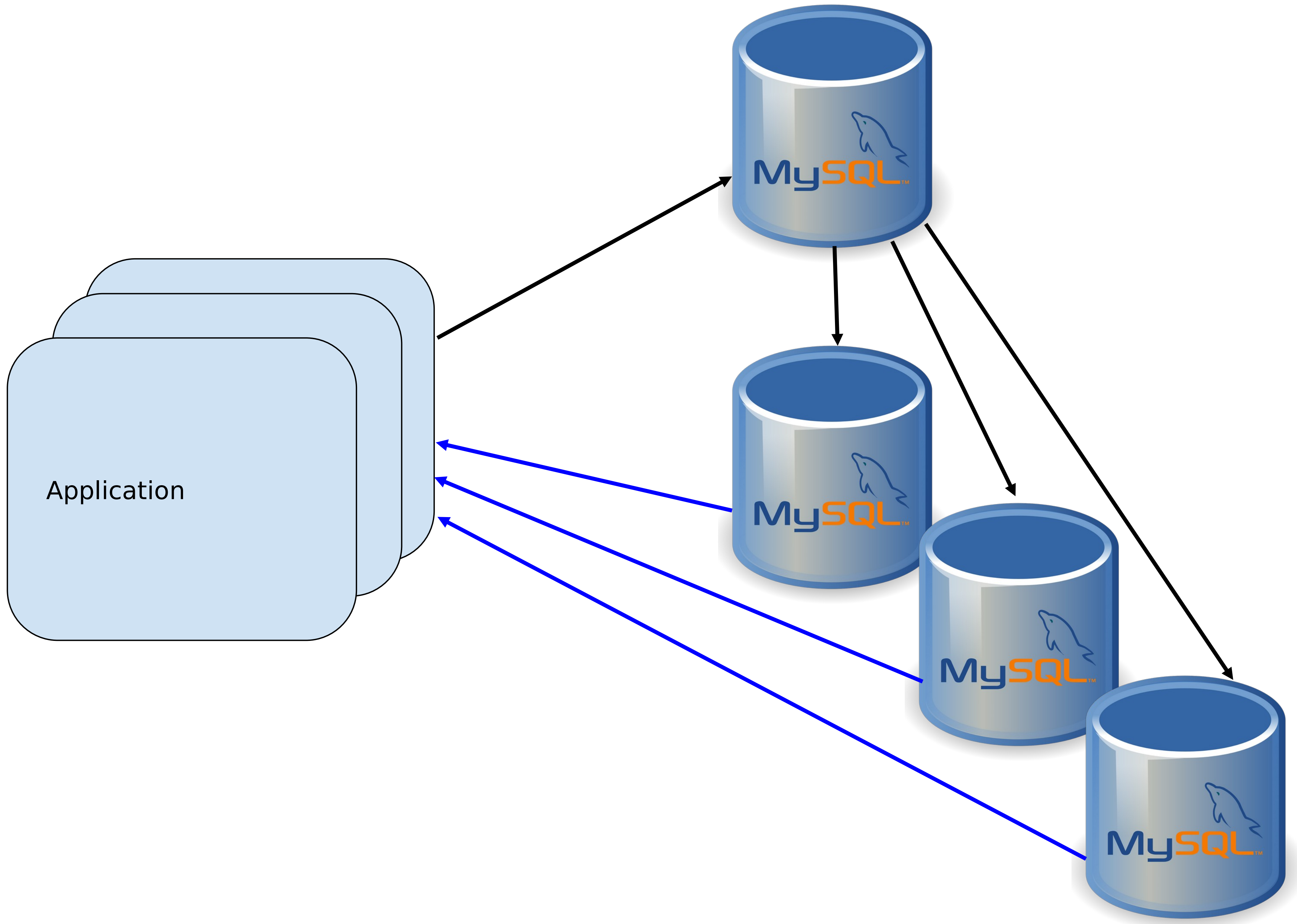
The problems with MySQL

MySQL was created around 1995
...for computer systems that were available at the time.
...for disks that were available at the time.
...for (inter)networks that were available at the time.

High availability with MySQL

An often-used setup with MySQL is to have a primary and multiple replicas.
Then if the primary fails you promote a replica.

Replica promotion is not automated.
Is your replica up to date?
Use a loadbalancer for service discovery?
How to fail back once the primary is back?
There is no leadership election.

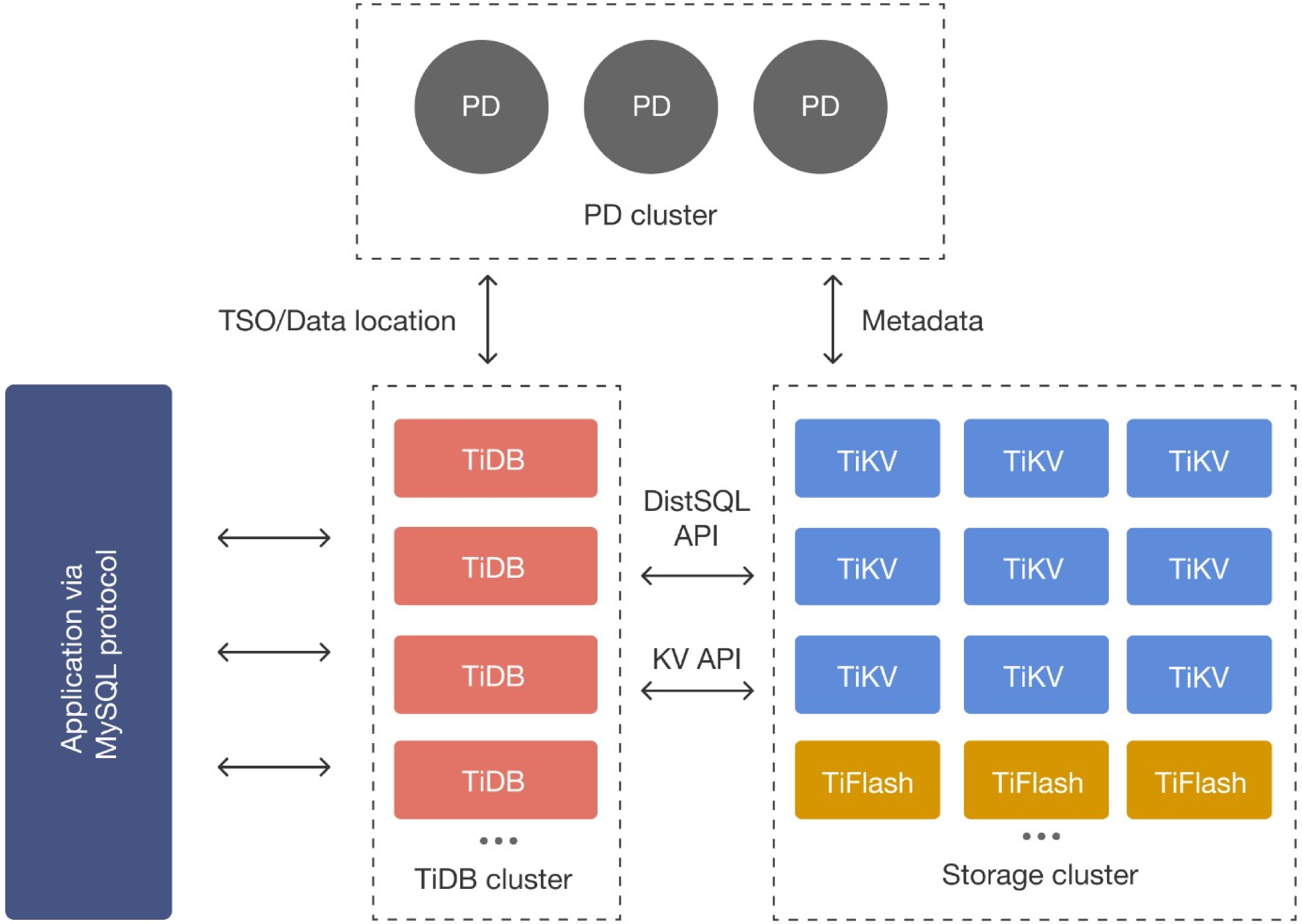


High availability with TiDB

All components are redundant.

The application is expected to retry failed connections.

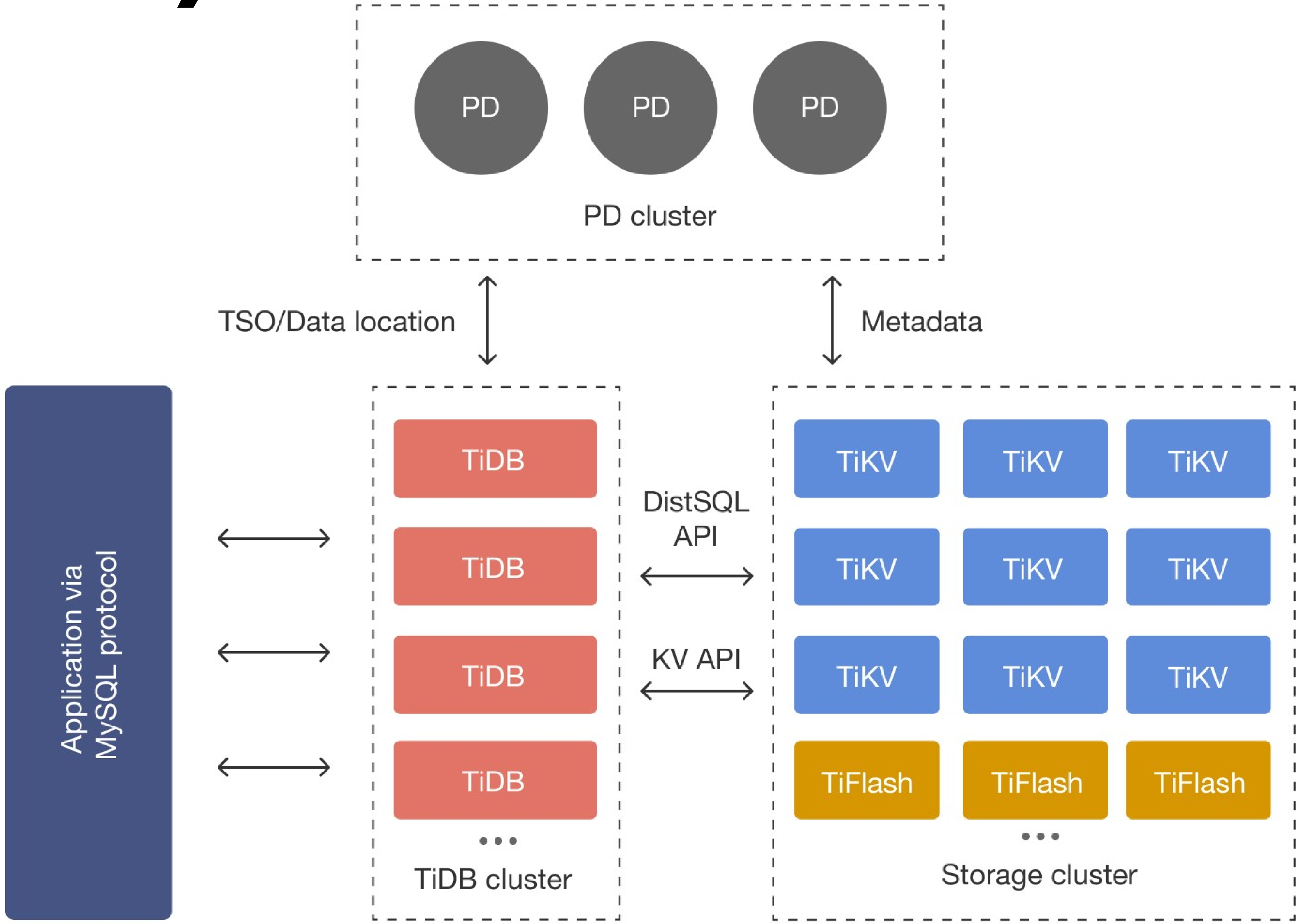
Tables are split into smaller parts of around 96 MiB each. These "data regions" are stored on TiKV.



Placement Driver (PD)

The placement driver:

- Is a Raft group itself
- Has etcd embedded
- Is the Time Stamp Oracle (TSO): it gives out timestamps. These timestamps are used in transactions and in the MVCC system
- Takes care of data placement:
  - TiKV servers have labels
  - labels are used to ensure a raft group spans multiple availability zones
  - Splitting big data regions, merging small data regions, splitting hot data regions, evenly distributing regions across the cluster.



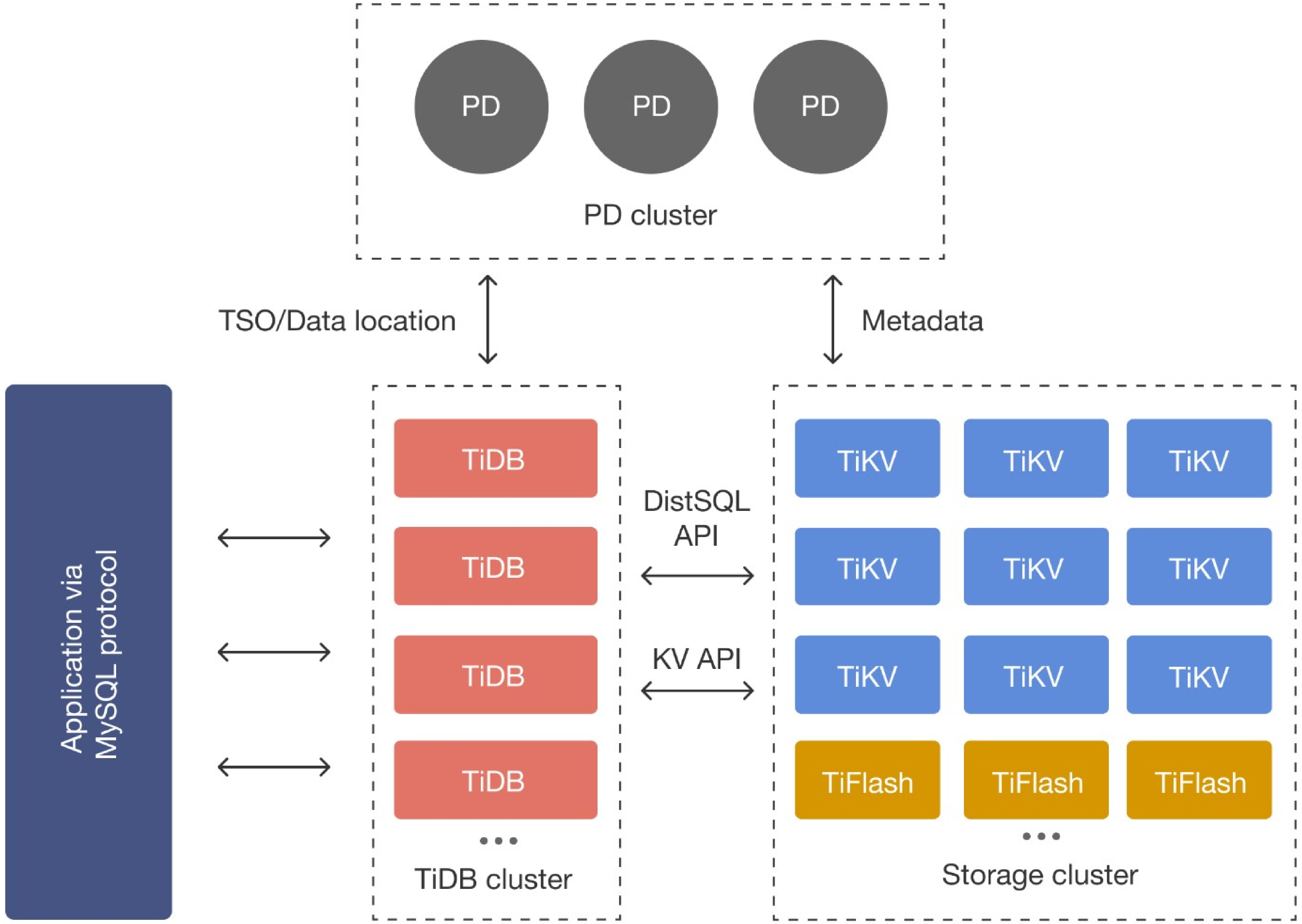
TiDB Server

TiDB Server is one of the components of the TiDB Platform. This is a bit confusing.

TiDB Server is written in Go and doesn't share any code with MySQL.

This implements the MySQL protocol and syntax.

TiDB is stateless, it doesn't store data.

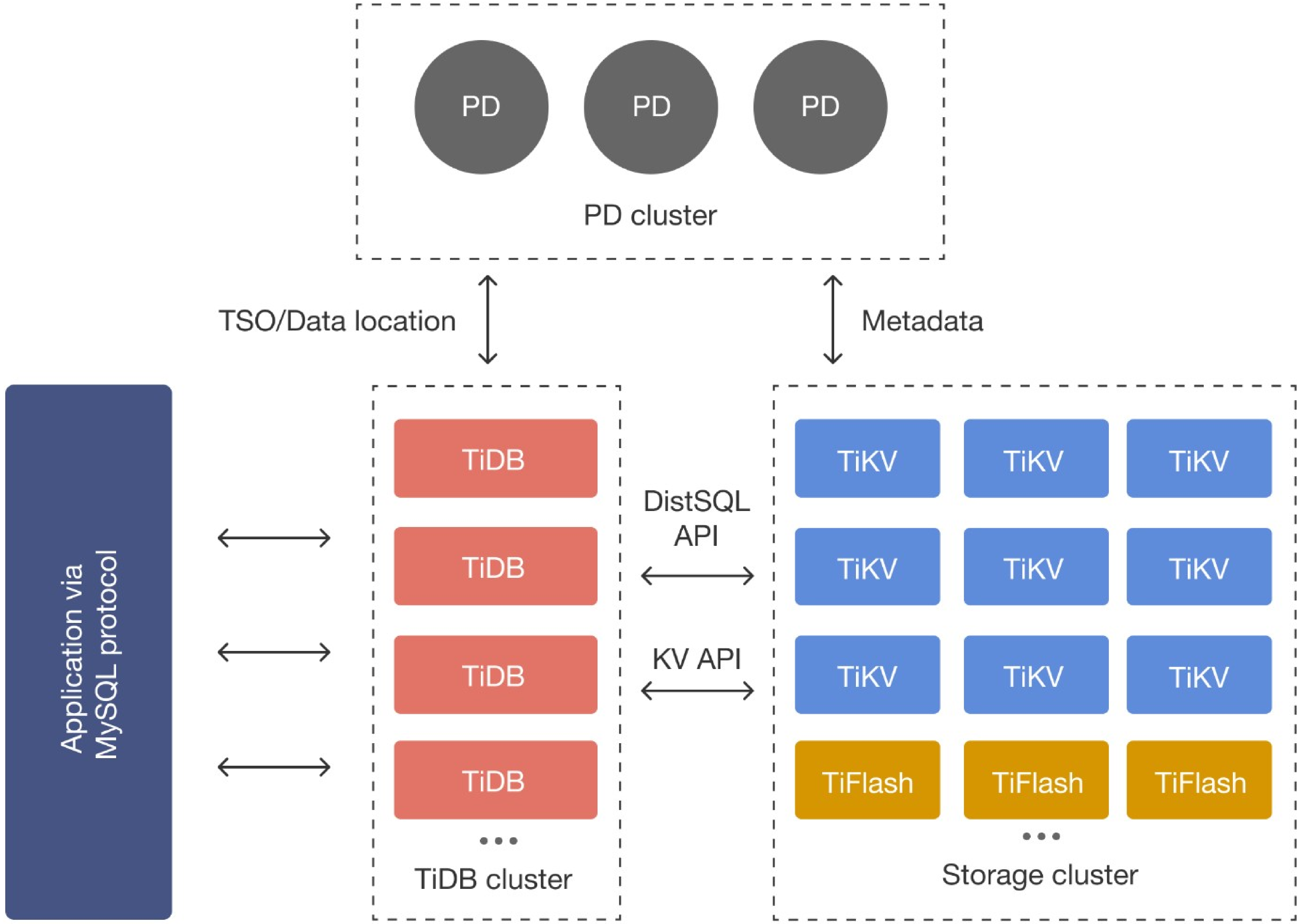


TiKV

TiKV is a key-value store. This is a CNCF project.

Database tables are stored with a RowID or PK as key and the columns as values.

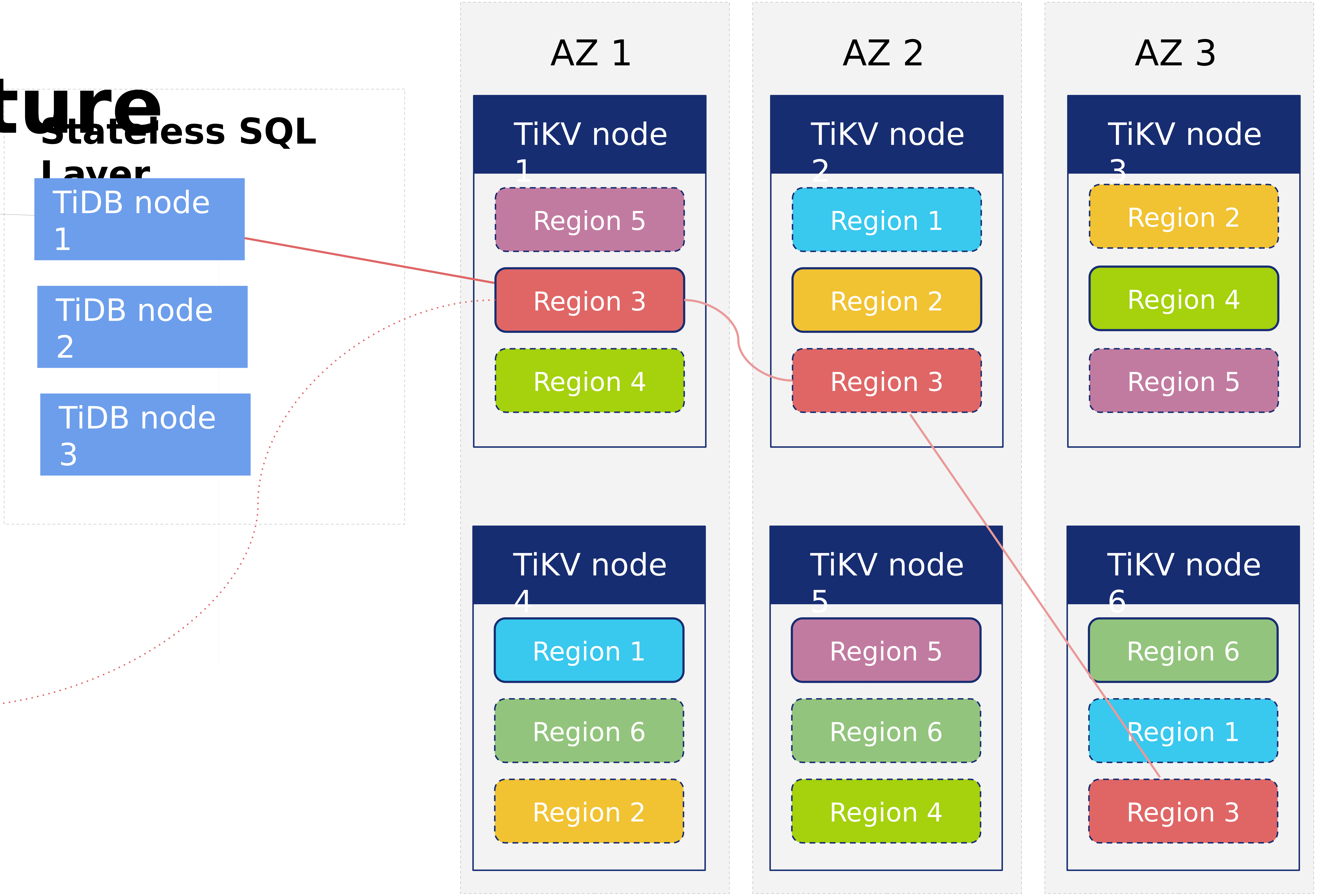
A database table is split up into multiple data regions. Each data region is a raft group of (by default) three nodes.



TiDB Architecture

SELECT id FROM orders WHERE id=1000001

orders	
1	data
2	data
…	…
1000001	…
…	…
999999000	…
…	…



Stateless SQL Layer

TiDB node 1

TiDB node 2

TiDB node 3

AZ 1

TiKV node 1

Region 5

Region 3

Region 4

TiKV node 4

Region 1

Region 6

Region 2

AZ 2

TiKV node 2

Region 1

Region 2

Region 3

TiKV node 5

Region 5

Region 6

Region 4

AZ 3

TiKV node 3

Region 2

Region 4

Region 5

TiKV node 6

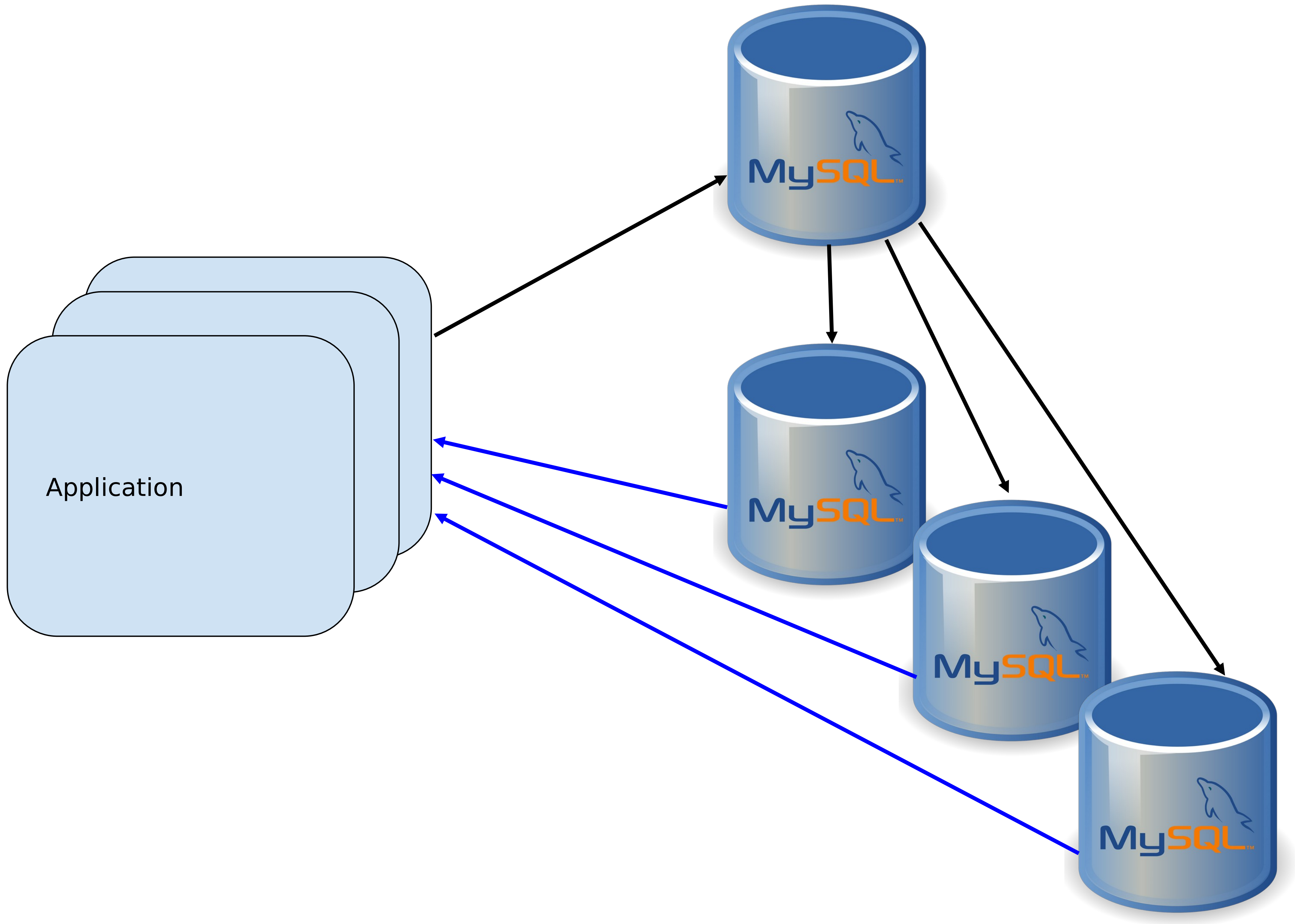
Region 6

Region 1

Region 3

Scalability with MySQL

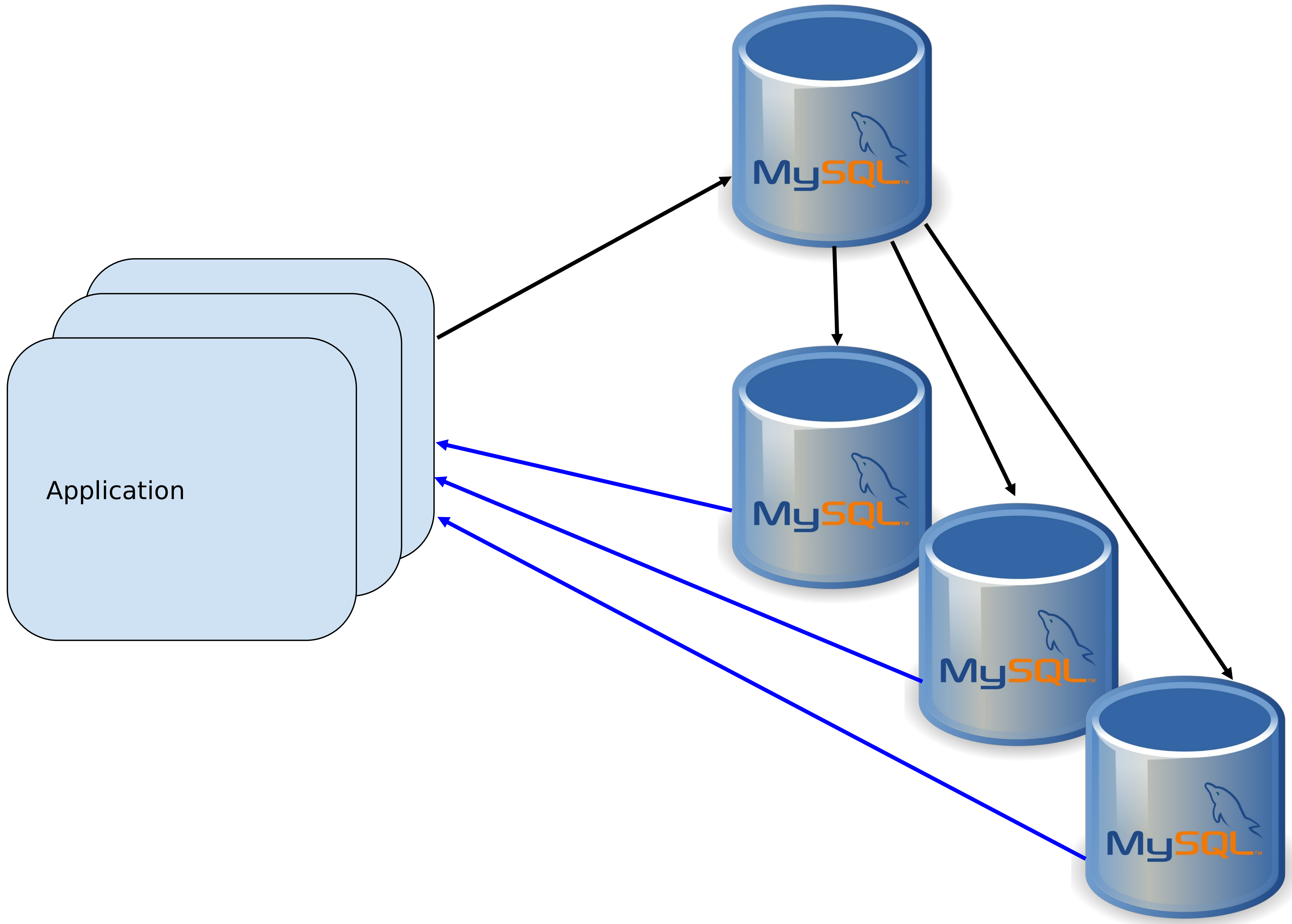
With a MySQL all writes go to the primary.
All nodes store a complete copy of the data.



Scalability with MySQL

Scaling reads?

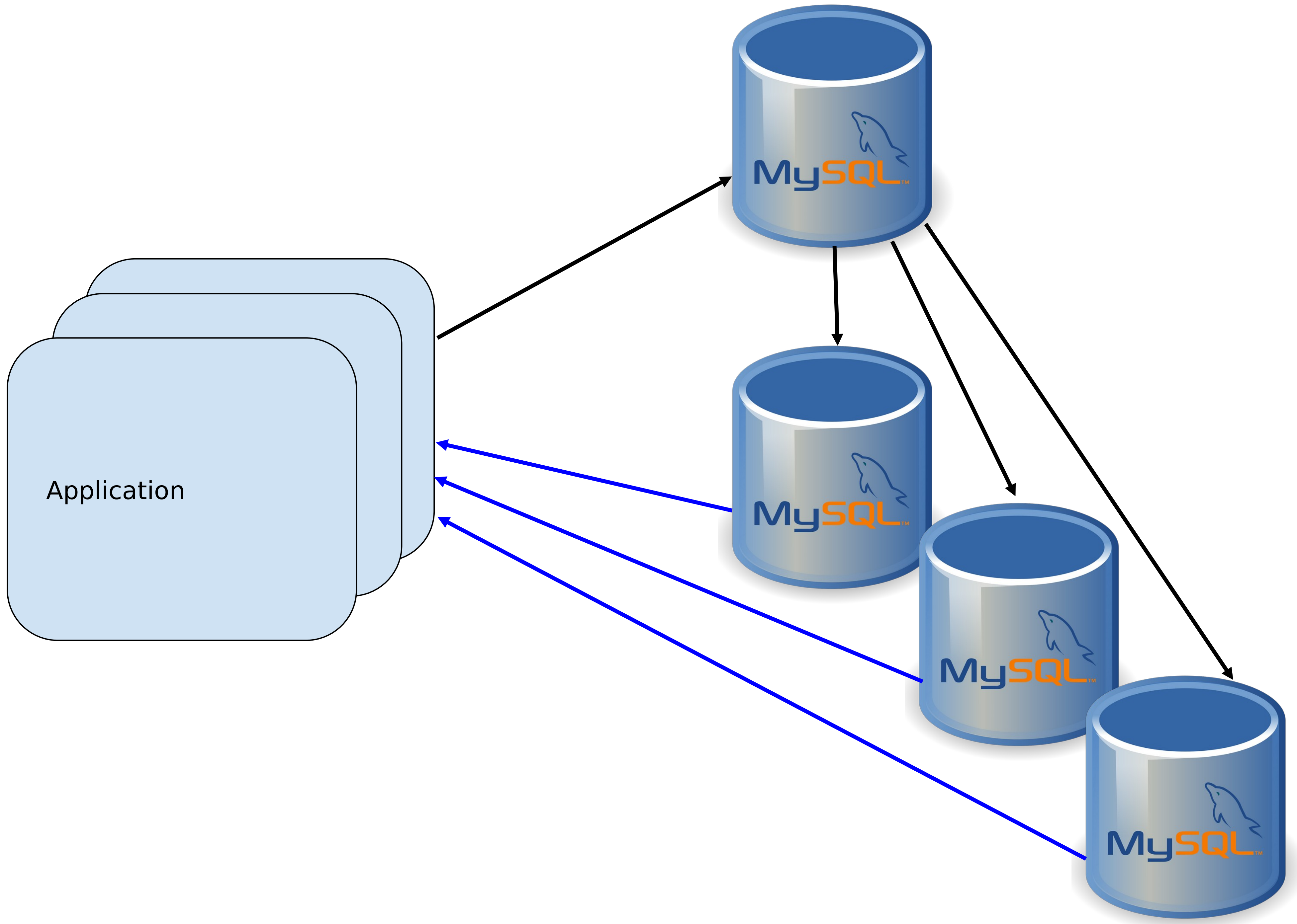
Add more replicas



Scalability with MySQL

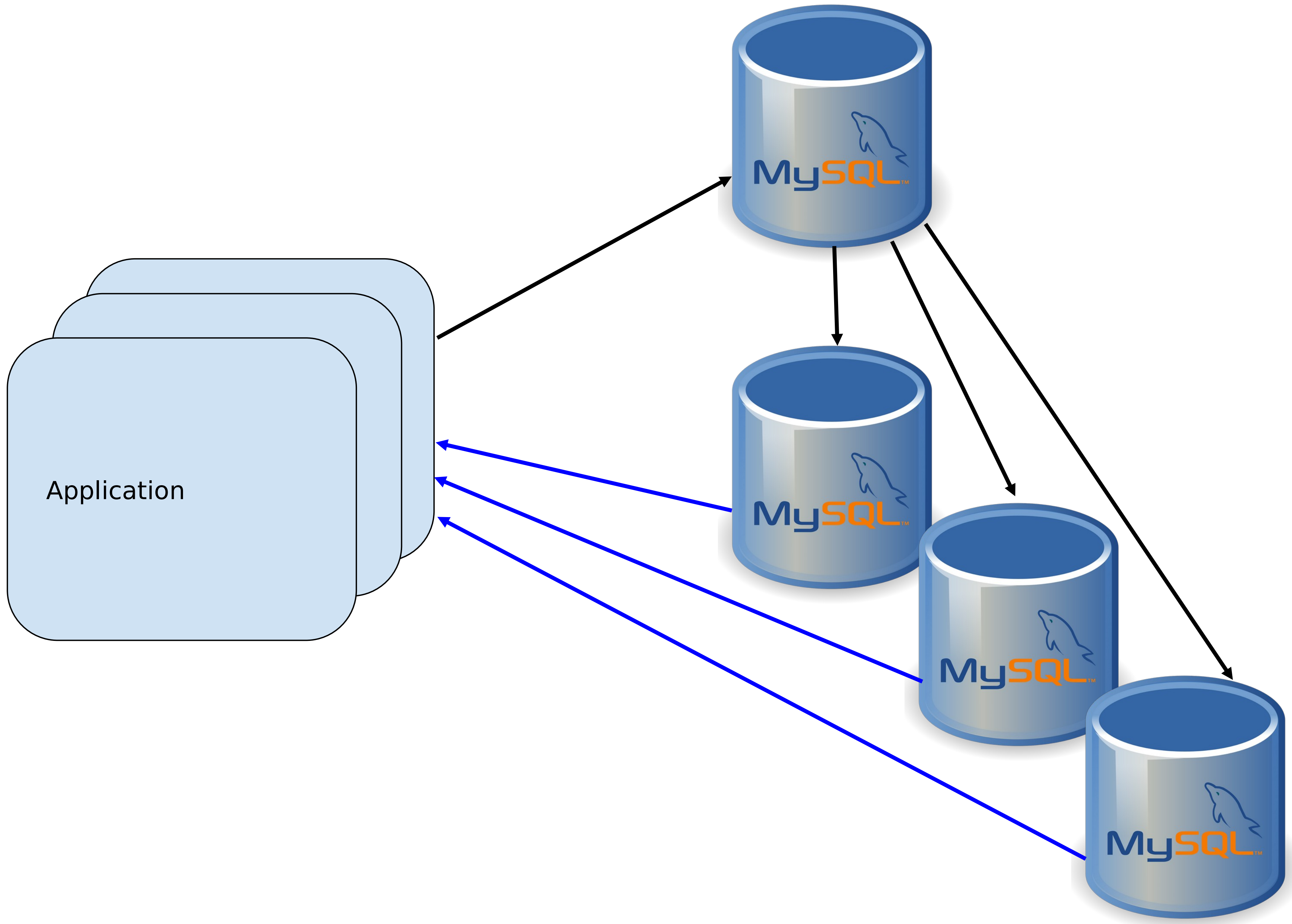
Scaling writes?

Replace the primary with a bigger machine



Scalability with MySQL

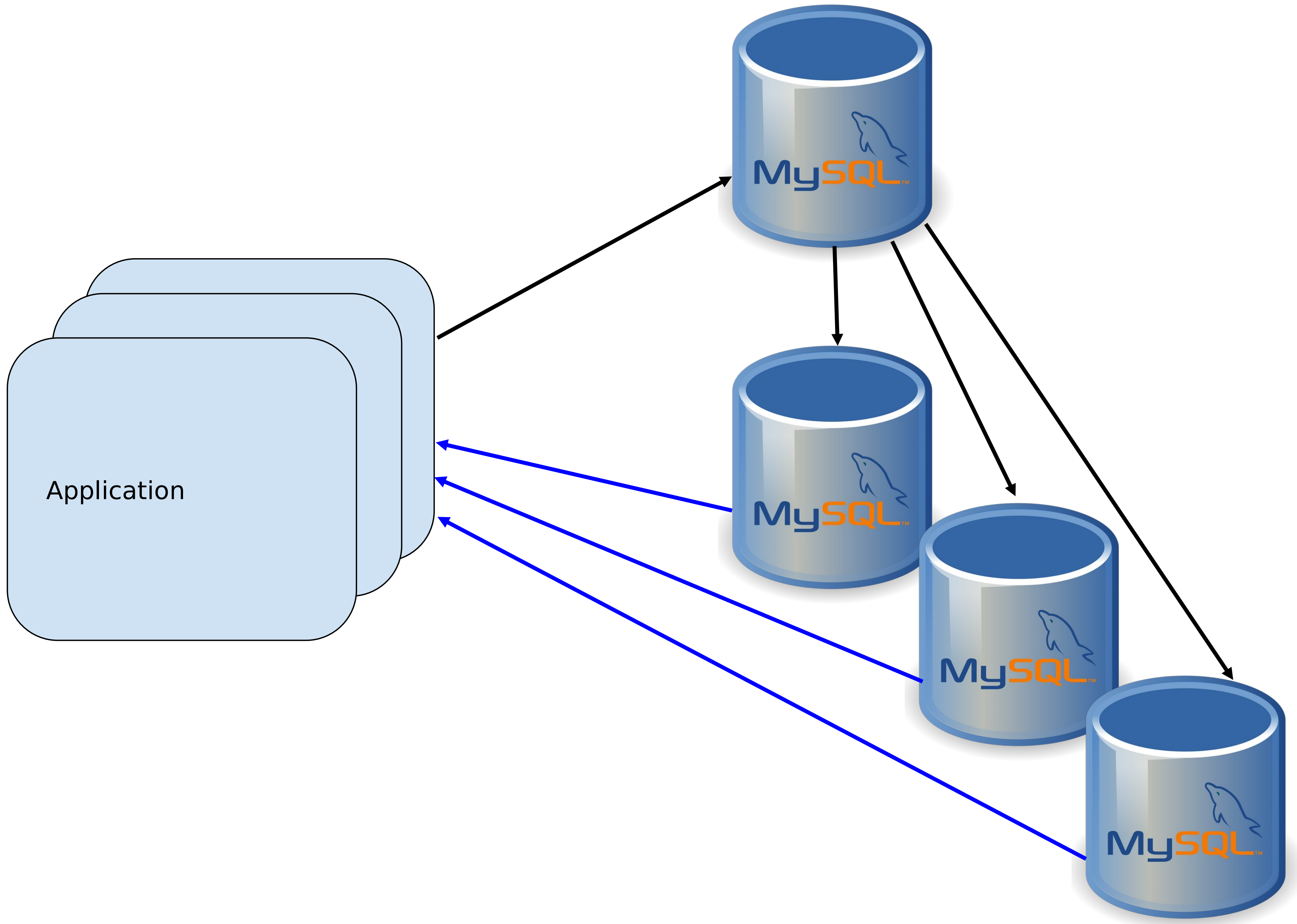
Scaling data volume?
Add bigger disks



Scalability with MySQL

Need to scale more?

Shard on the application side



Scalability with TiDB

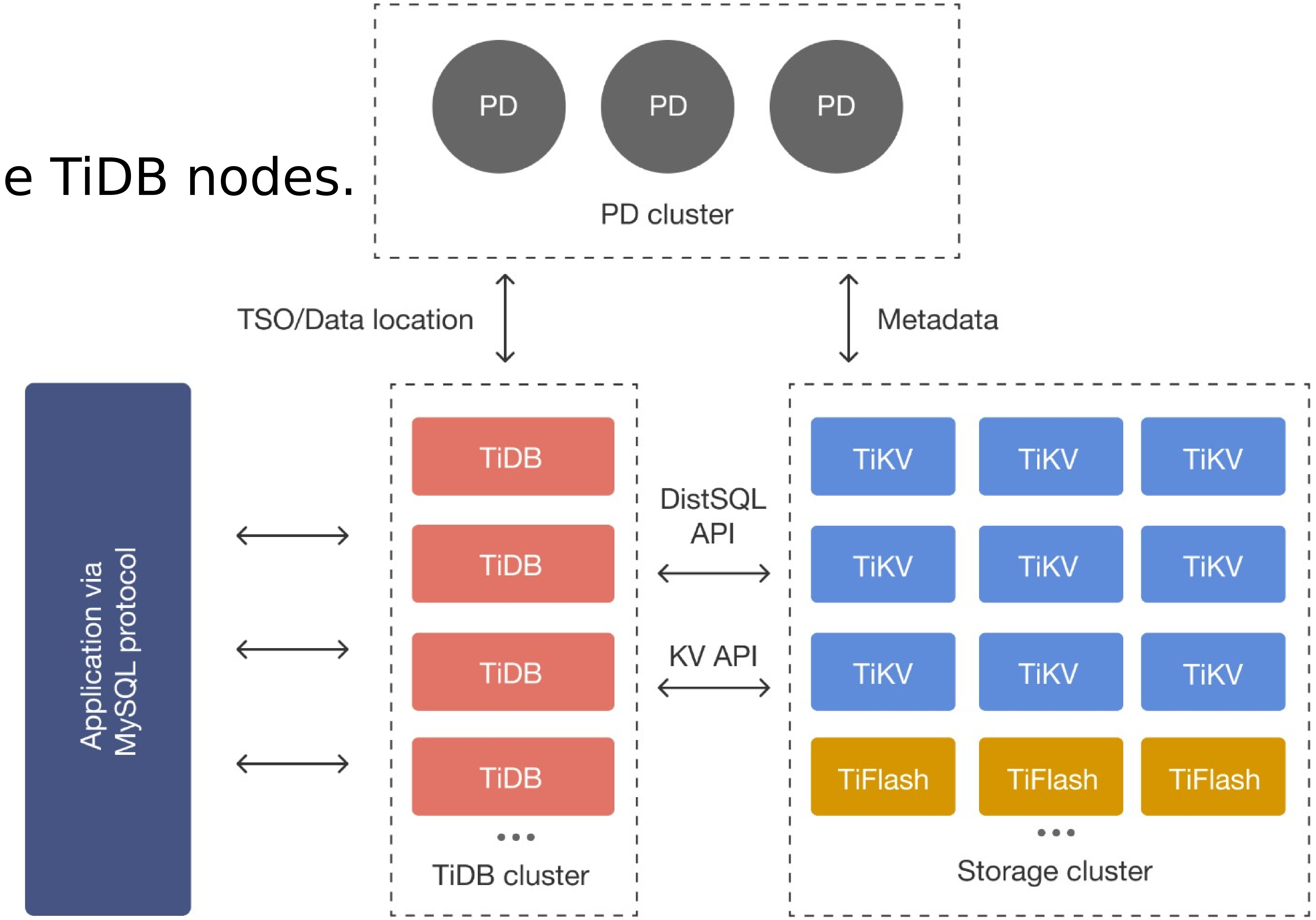
Both reads and writes can go to any of the TiDB nodes.

Scaling reads?
    Add more nodes

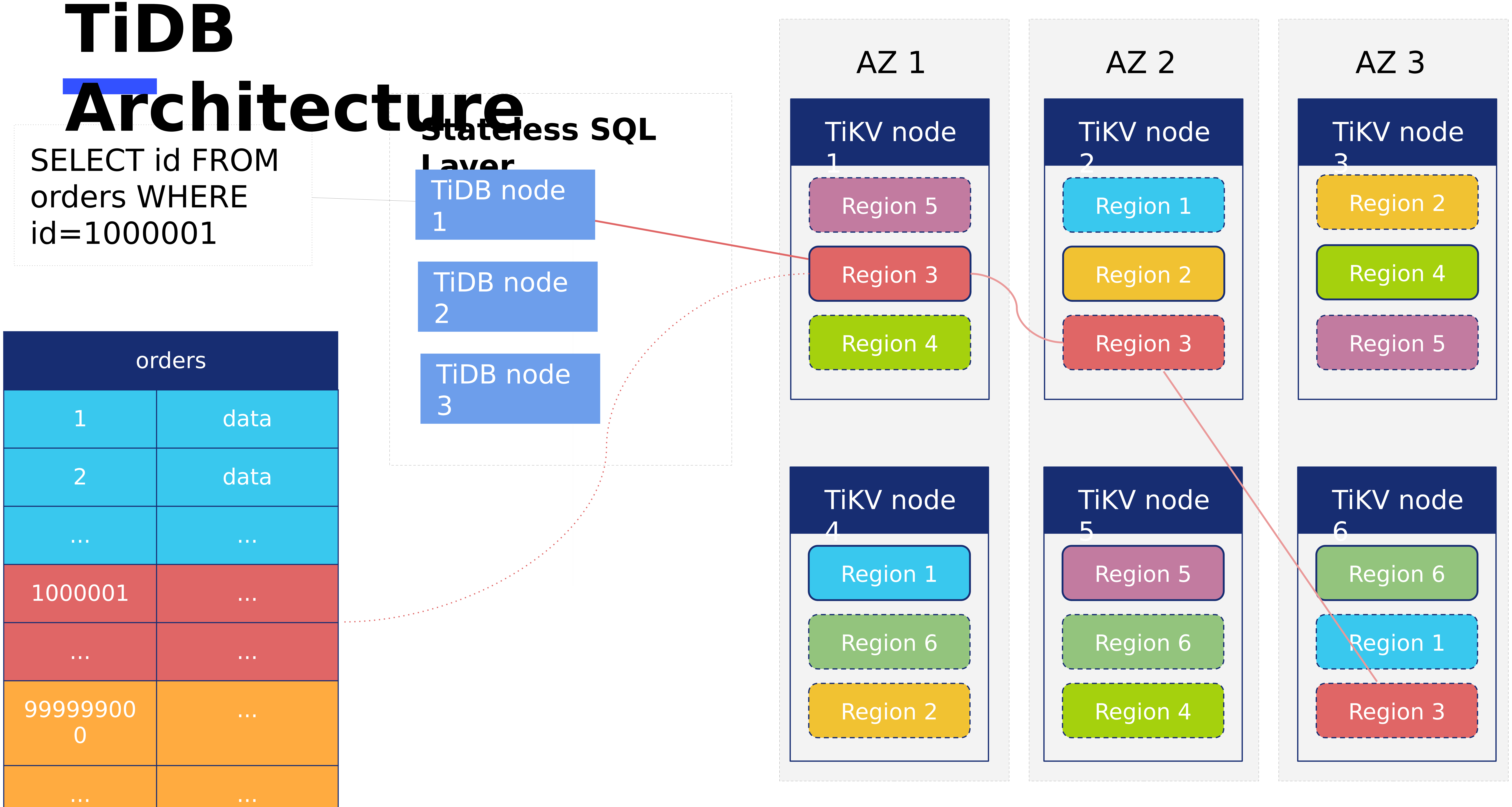
Scaling writes?
    Add more nodes

Scaling data volume?
    Add more nodes

Need to scale more?
    Add more nodes



TiDB Architecture



SELECT id FROM orders WHERE id=1000001

Stateless SQL Layer

TiDB node 1

TiDB node 2

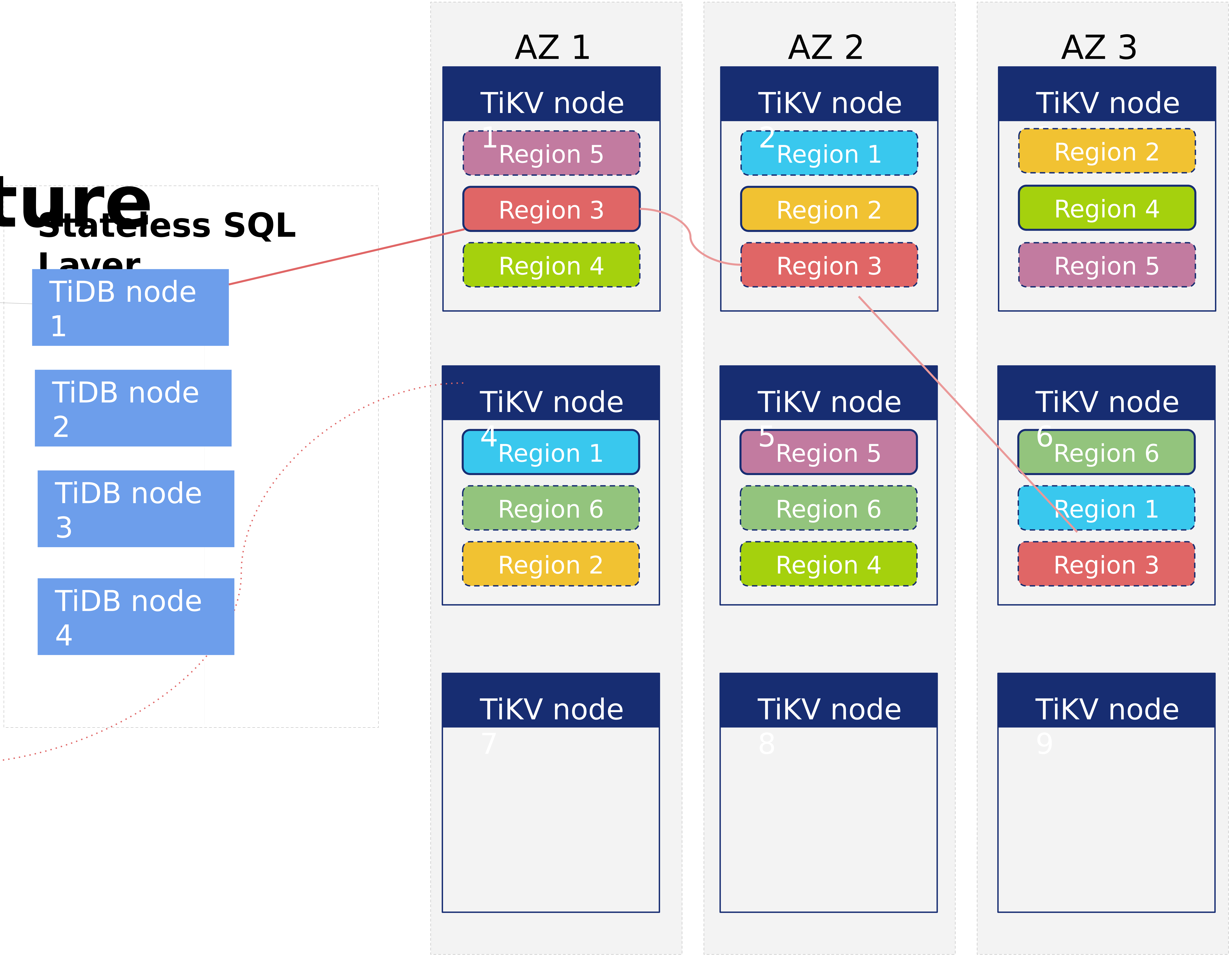
TiDB node 3

orders	
1	data
2	data
…	…
1000001	…
…	…
99999900 0	…
…	…

TiDB Architecture

SELECT id FROM orders WHERE id=1000001

orders	
1	data
2	data
…	…
1000001	…
…	…
999999000	…
…	…



Stateless SQL Layer

TiDB node 1

TiDB node 2

TiDB node 3

TiDB node 4

AZ 1

TiKV node 1
Region 5
Region 3
Region 4

TiKV node 4
Region 1
Region 6
Region 2

TiKV node 7

AZ 2

TiKV node 2
Region 1
Region 2
Region 3

TiKV node 5
Region 5
Region 6
Region 4

TiKV node 8

AZ 3

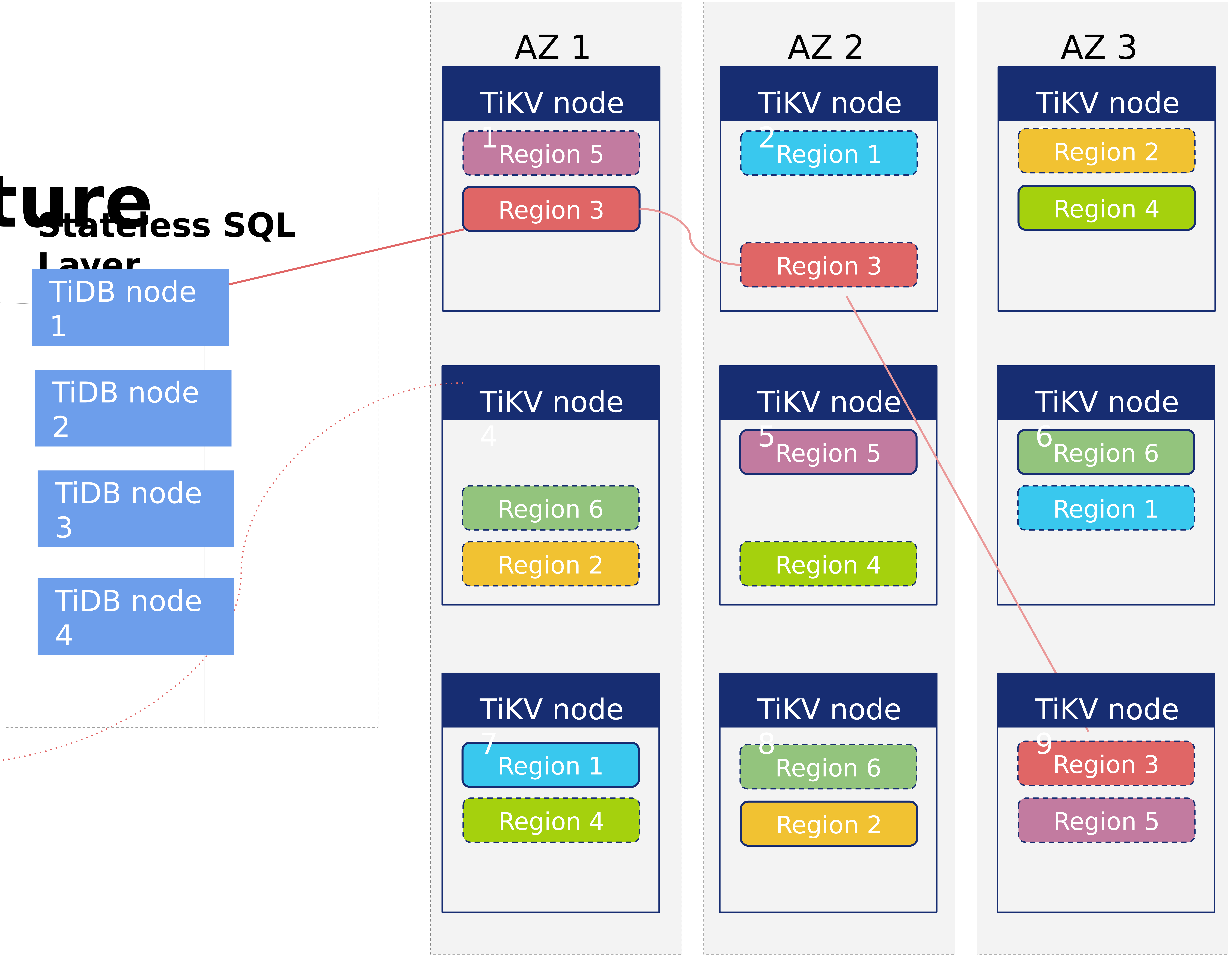
TiKV node
Region 2
Region 4
Region 5

TiKV node 6
Region 6
Region 1
Region 3

TiKV node 9

TiDB Architecture

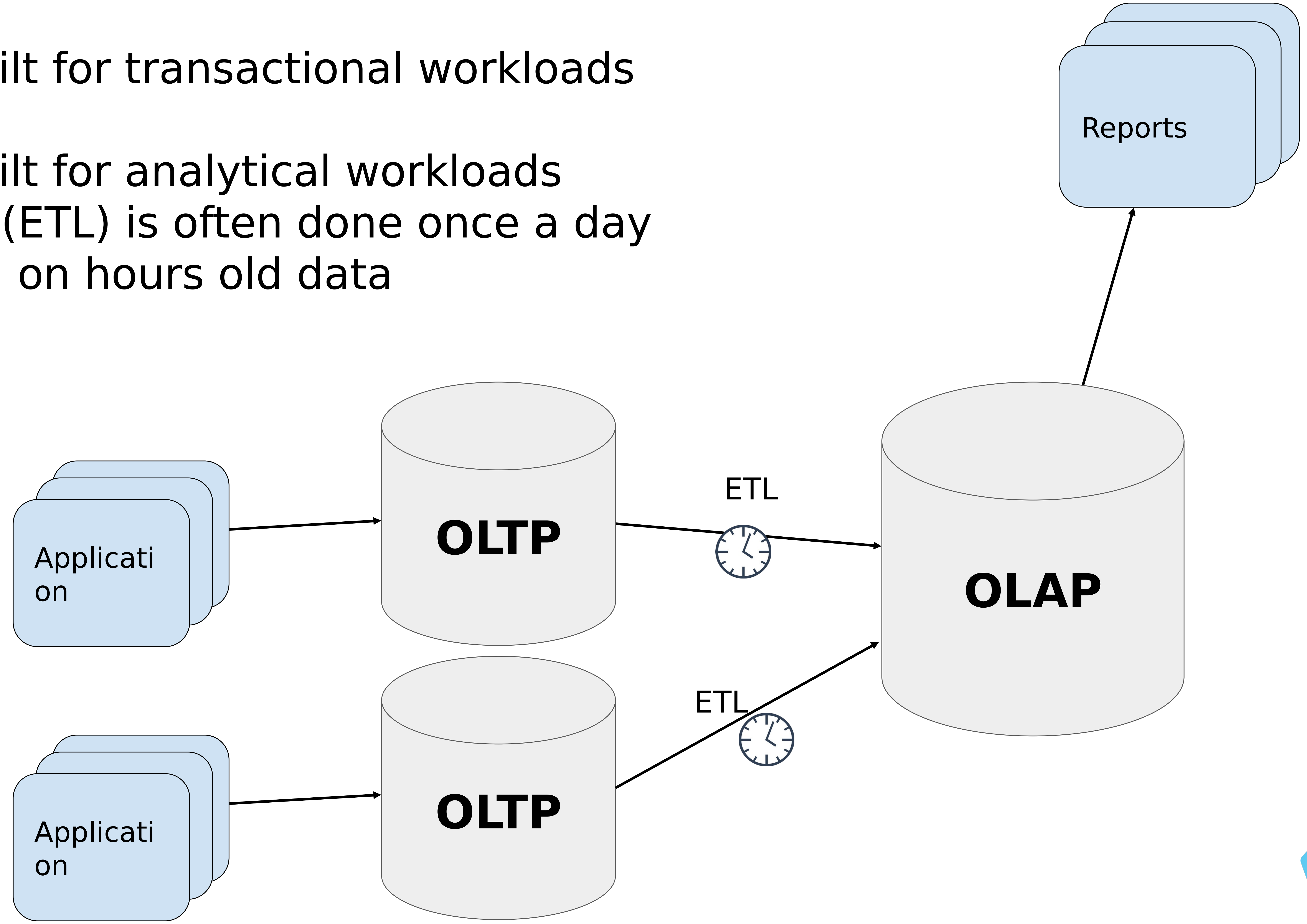
SELECT id FROM orders WHERE id=1000001



orders	
1	data
2	data
…	…
1000001	…
…	…
999999000	…
…	…

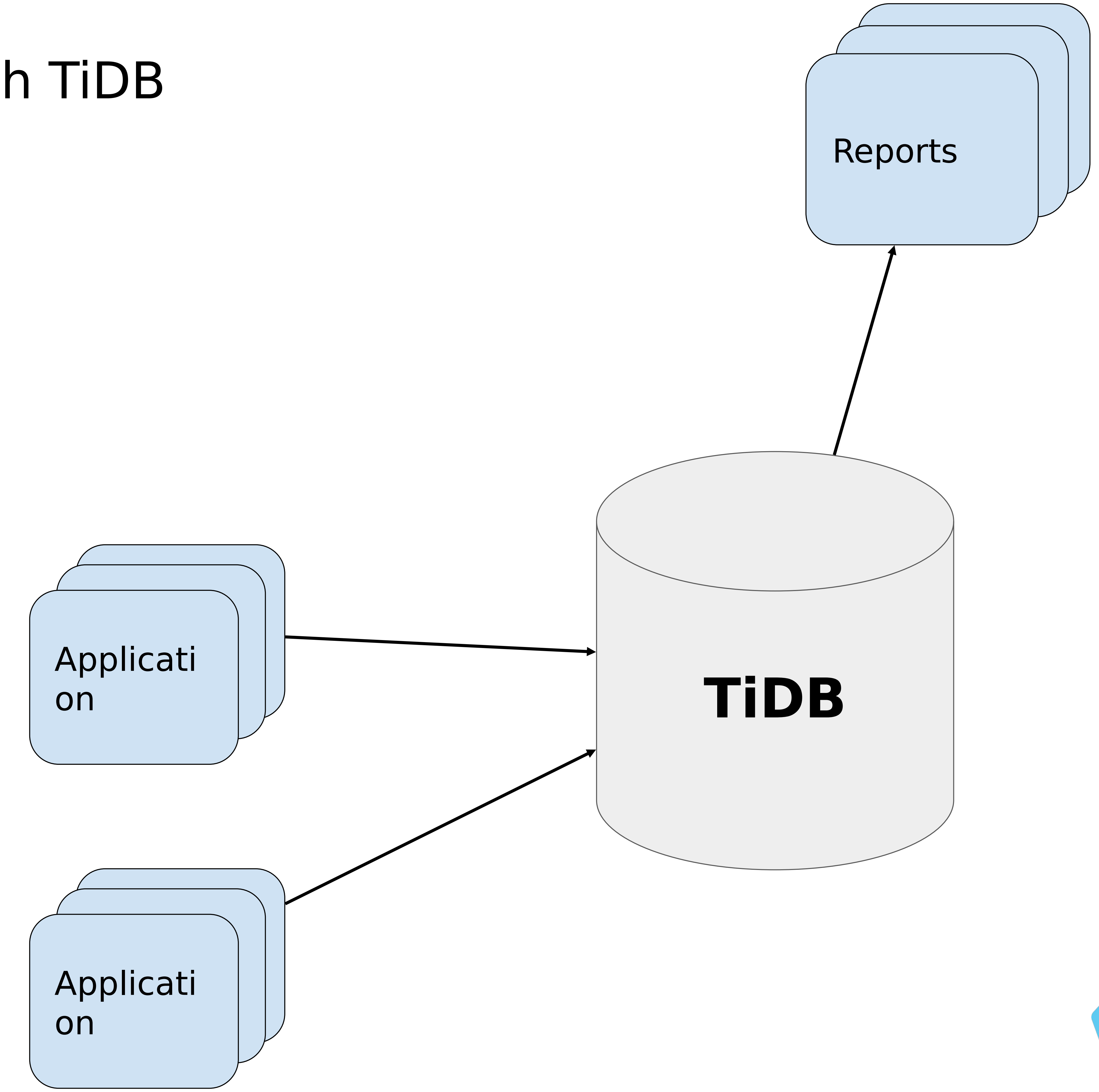
Analytics and ETL

- Some databases are built for transactional workloads (OLTP)
- Some databases are built for analytical workloads
- Extract-Transform-Load (ETL) is often done once a day
- Reports might be based on hours old data



Analytics and ETL

- Replace both OLTP and OLAP systems with TiDB
- No more ETL required
- No more outdated reports



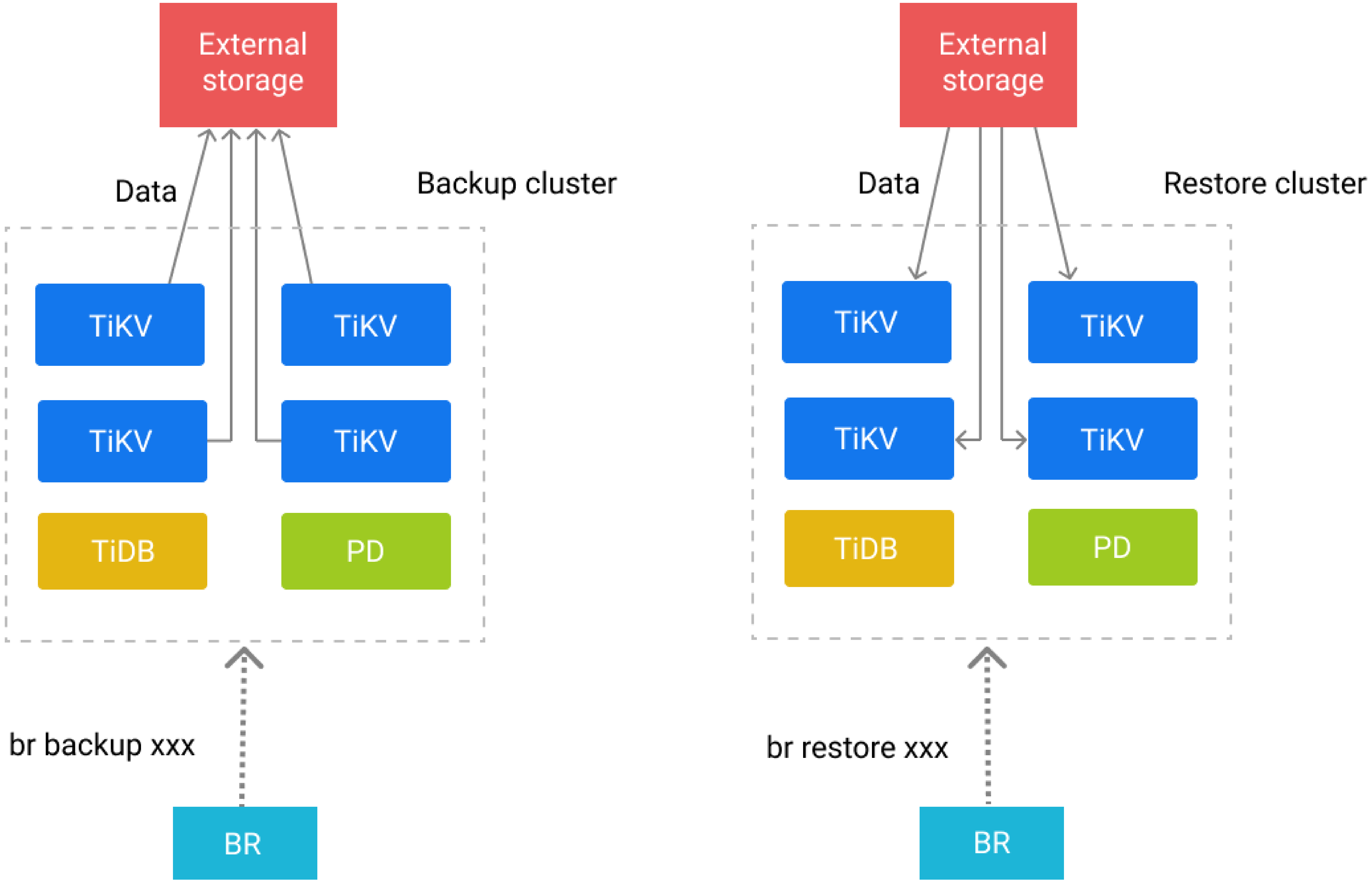
Analytics and ETL

- TiFlash stores **one or more copies** of selected tables in row based format. The primary copy is always on TiKV.
- TiFlash is based on Clickhouse.
- TiFlash joins the raft groups as a learner. It will never become a leader.
- To use TiFlash you set the number of copies on a per table level. Multiple copies are good for redundancy and allow parallel execution of queries.
- The optimizer is smart enough to select the row store (TiKV) or the column store (TiFlash) based on the type of query.
- Execution of a query can use both TiKV and TiFlash for executing different parts of the query.
- Transaction isolation is guaranteed, also when TiFlash is used.

Batteries included: BR

Backup and Restore (BR)

- The TiKV servers all write to shared storage to create a backup.
- The TiKV servers all read from shared storage to restore.
- This makes backups fast and scalable.
- External storage could be S3, MinIO, GCS or a shared filesystem.
- Opensource.



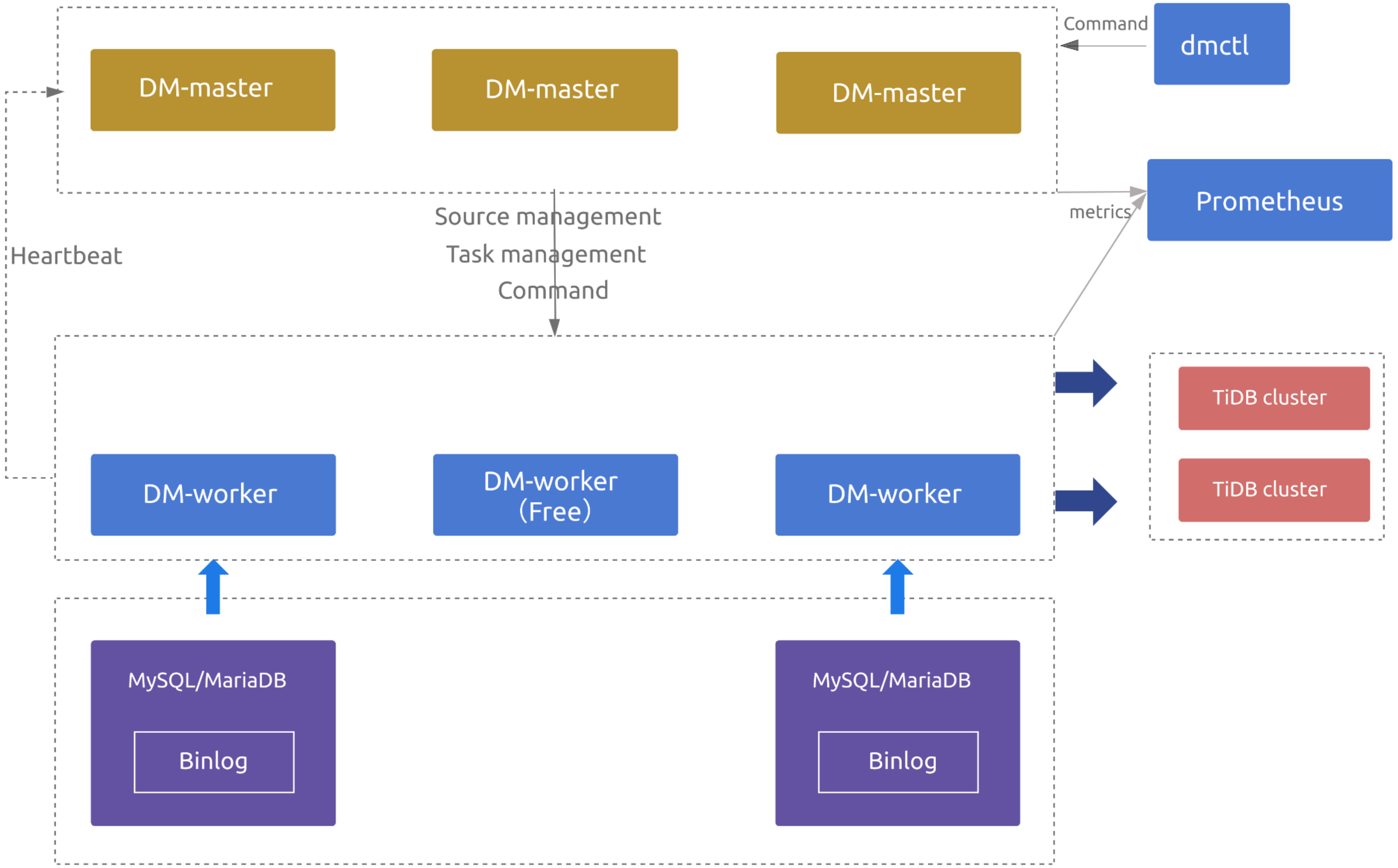
Batteries included

- Dumpling
  - Dumping data to SQL or CSV in parallel. Can also dump data from MySQL.
- Lightning
  - Restore a dump that was made with dumpling.
  - Can import CSV made by other tools.
  - Can import directly to TiKV, bypassing the SQL layer.
  - Can also import via SQL statements to TiDB.
- Main usecase is to migrate data from MySQL

Batteries included

Data Migration (DM)

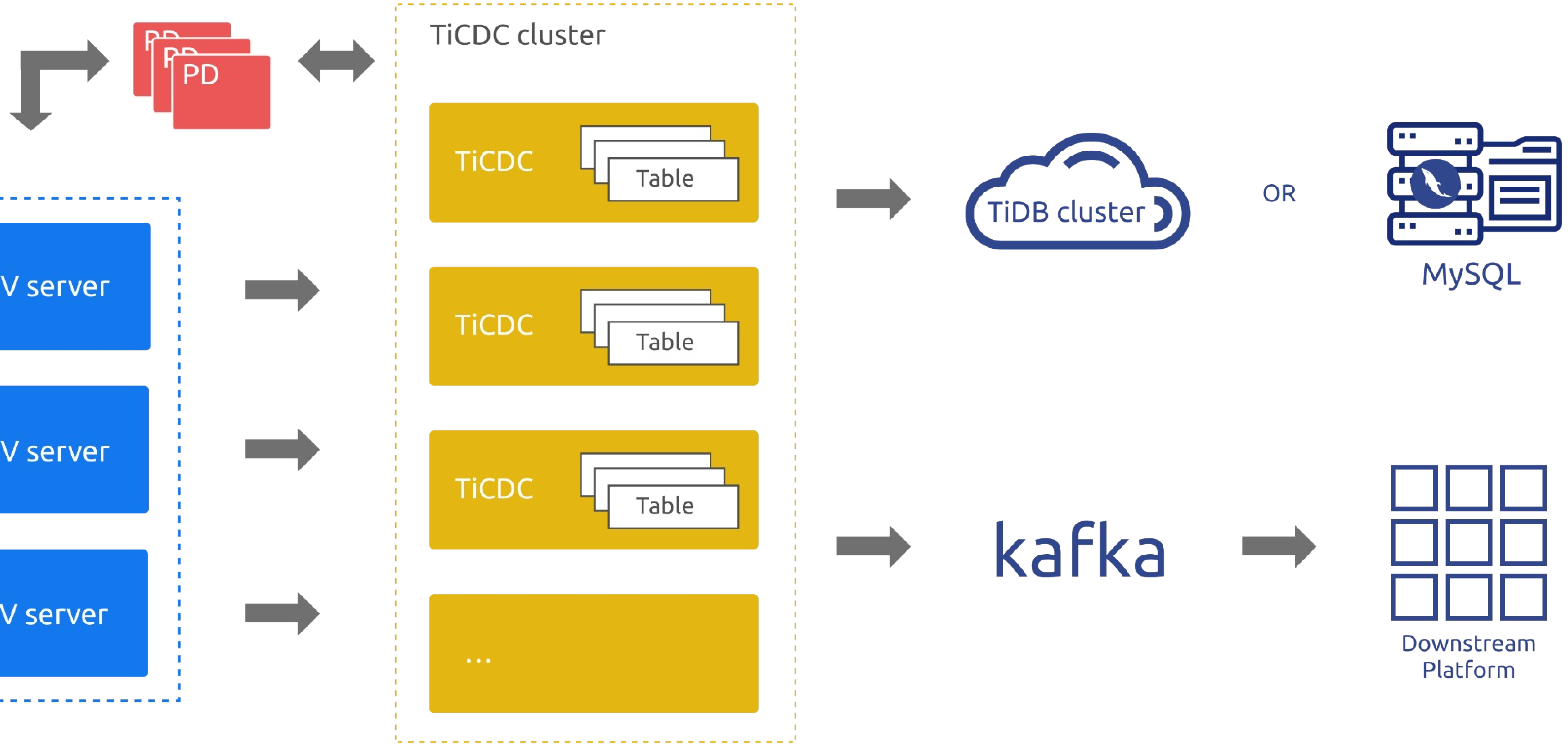
- Replicate data from MySQL to TiDB.
- Uses Dumpling/Lightning to copy the initial copy.
- Note that this is also made to be high available.



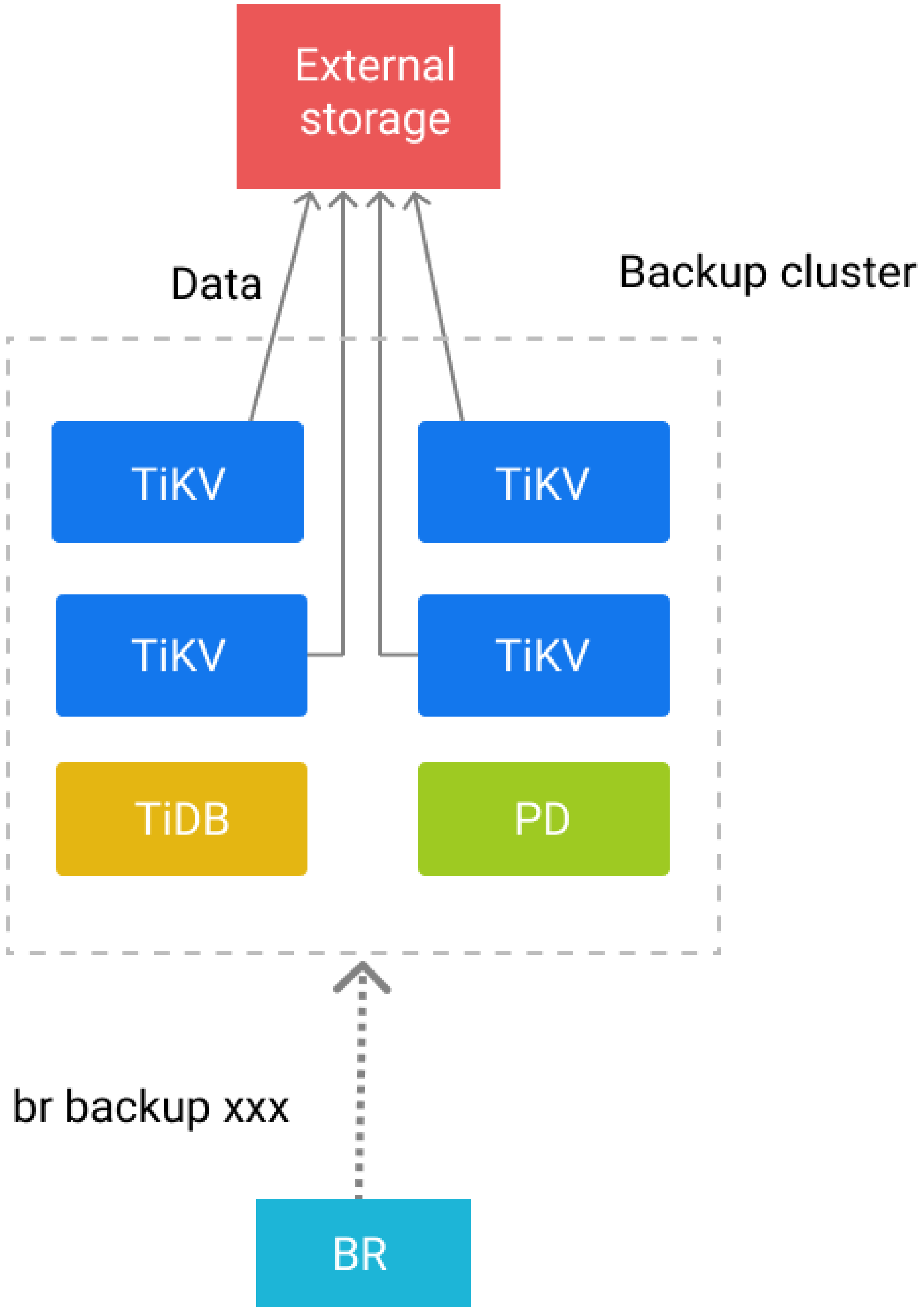
Batteries included

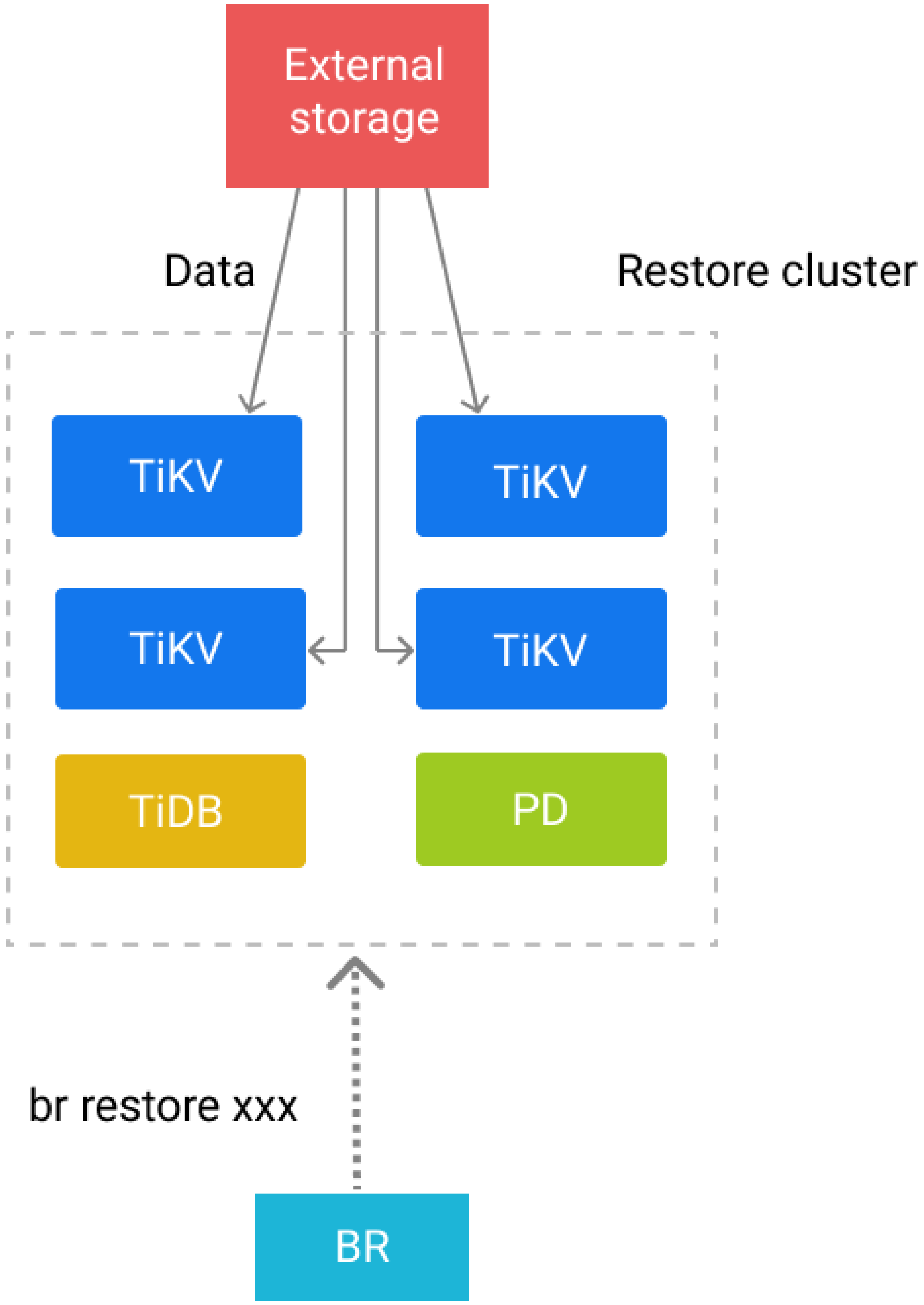
TiCDC

- Change Data Capture
- Send events to Kafka, MySQL or another TiDB Cluster.
- Also high available



BR


External storage
Data
Backup cluster
TiKV
TiKV
TiKV
TiKV
TiDB
PD
br backup xxx
BR


External storage
Data
Restore cluster
TiKV
TiKV
TiKV
TiKV
TiDB
PD
br restore xxx
BR

Leave your feedback!

You can rate the talk and give feedback on what you've liked or what could be improved

я

++ High Load Armenia

Co-organizer

Yandex